Úkol C – Klasifikace dat – Spambase

ZS 2016

Cíl

Cílem třetího úkolu je seznámit se s praktickým použitím klasifikačních metod na příkladě predikce klasifikace elektronické pošty na normální a nevyžádanou poštu.

Data

Detekce spamu je jeden z typických problémů strojového učení. Nevyžádaná pošta často slouží k reklamě, k šíření škodlivého softwaru (viry), získávání osobních informací (phishing) a nebo k jednoznačným podvodům (Nigerijský princ). Zpracováním emailů pro potřeby strojového učení se zabývá text mining, se kterým se setkáme na některém z následujících cvičení. V případě tohoto úkolu jsou už data upravena do formy datové matice, a není proto třeba provádět jakékoli zpracování skutečných emailů.

Popis dat

Data obsahují informace o 4600 emailech. Emaily jsou popsány pomocí výskytu některých slov a znaků (money, credit, $, ...), které jsou v proměnných ‚word_freq_' a ‚char_freq_' jako procento z celkového počtu slov/znaků. Dalšími proměnnými jsou počty velkých písmen, ‚capital_run_length_average' je průměrná délka řetězců slov uvedených ve velkých písmenech, ‚capital_run_length_longest' je největší délka řetězce slov uvedených ve velkých písmenech a ‚capital_run_length_total' je celková délka řetězců slov uvedených ve velkých písmenech.

Klasifikace elektronické pošty

Pokuste se vytvořit takový na základě předložených dat klasifikátor, který by predikoval, zda email je a nebo není nevyžádaná pošta. Klasifikátor zhodnoťte z různých praktických i teoretických hledisek.

Požadované kroky analýzy

- Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat? [2 b]
- Na vhodné podmnožině dat vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? [2 b]
- Podle jakých příznaků se klasifikátor rozhoduje? Dává to smysl? Lze na základě vaší analýzy omezit počet měřených příznaků při zachování stejné úspěšnosti klasifikace? [4 b]
- Jakou úspěšnost klasifikace očekáváte v hypotetickém reálném nasazení vašeho klasifikátoru, tj. v případě nově příchozí pošty? [4 b]
- Jaká je pravděpodobnost na základě dat, že nově příchozí pošta je nevyžádaná? [2 b]
- Jaká je pravděpodobnost, že nově příchozí pošta bude klasifikována jako nevyžádaná? Jaká bude naproti tomu pravděpodobnost, že nově příchozí pošta nebude klasifikována jako nevyžádaná. Výsledky diskutujte. [4 b]
- Má ve vašem případě na přesnost klasifikace vliv to, zda je trénovací (testovací) množina vyvážená? Pokud ano, jaký? Vyvažovali jste trénovací (testovací) množinu? Pokud ano, proč a jak? Pokud ne, proč ne? [2 b]

Výsledky upravte do formy zprávy, která bude obsahovat stručný **úvod**, popis metod, které jste použili, v sekci **metody,** výsledky jejich aplikace na data v sekci **výsledky** a závěry, které jste zjistili interpretací výsledků v sekci **závěr**. Maximální délka zprávy je 3 stránky. Zprávu ve formátu pdf odevzdejte pomocí UploadSystemu.